

(12) NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES
PATENTWESENS (PCT) VERÖFFENTLICHTE INTERNATIONALE ANMELDUNG

(19) Weltorganisation für geistiges Eigentum
Internationales Büro

(43) Internationales Veröffentlichungsdatum
04. Juni 2020 (04.06.2020)



(10) Internationale Veröffentlichungsnummer
WO 2020/109016 A1

(51) Internationale Patentklassifikation:

G06K 9/00 (2006.01) G06K 9/62 (2006.01)
G06K 9/32 (2006.01) G06T 7/246 (2017.01)

(21) Internationales Aktenzeichen: PCT/EP2019/081317

(22) Internationales Anmeldedatum:

14. November 2019 (14.11.2019)

(25) Einreichungssprache: Deutsch

(26) Veröffentlichungssprache: Deutsch

(30) Angaben zur Priorität:

10 2018 220 276.1

26. November 2018 (26.11.2018) DE

10 2018 220 274.5

26. November 2018 (26.11.2018) DE

(71) Anmelder: **OSRAM GMBH** [DE/DE]; Marcel-Breuer-Straße 6, 80807 München (DE).

(72) Erfinder: **AMIN, Sikandar**; Heideweg 2, 85375 Neufahrn B. Freising (DE). **MUNJAL, Bharti**; Stuntzstraße 17, 81677 München (DE). **BRANDLMAIER, Meltem Demirkus**; Armulfstr. 65, 80636 München (DE). **AFTAB, Abdul Rafey**; Grünwalderstrasse 9, 81547 München (DE). **GALASSO, Fabio**; Münchener Str. 24D, 85748 Garching (DE).

(81) Bestimmungsstaaten (soweit nicht anders angegeben, für jede verfügbare nationale Schutzrechtsart): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Bestimmungsstaaten (soweit nicht anders angegeben, für jede verfügbare regionale Schutzrechtsart): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), eurasisches (AM, AZ, BY, KG, KZ,

RU, TJ, TM), europäisches (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Veröffentlicht:

— mit internationalem Recherchenbericht (Artikel 21 Absatz 3)

(54) Title: METHOD FOR COMMON DETECTING, TRACKING AND CLASSIFYING OF OBJECTS

(54) Bezeichnung: VERFAHREN ZUM GEMEINSAMEN DETEKTIEREN, VERFOLGEN UND KLASSIFIZIEREN VON OBJEKTEN

(57) Abstract: The invention relates to a method for machine-based training of a computer-implemented network for common detecting, tracking and classifying of at least one object in a video image sequence having a plurality of successive individual images. A combined error is determined during the training, which error results from the errors of the determining of the class identification vector, the determining of the at least one identification vector, the determining of the specific bounding box regression, and the determining of the inter-frame regression.

(57) Zusammenfassung: Es wird ein Verfahren für ein maschinenbasiertes Training eines computerimplementierten Netzwerks zum gemeinsamen Detektieren, Verfolgen und Klassifizieren wenigstens eines Objektes in einer Videobildsequenz aufweisend eine Vielzahl aufeinanderfolgender Einzelbilder vorgeschlagen. Dabei wird während des Trainings ein kombinierter Fehler ermittelt, der sich aus den Fehlern des Bestimmens des Klassenidentifikationsvektors, des Bestimmens des wenigstens eines Identifikationsvektors, des Bestimmens der spezifischen bounding box regression und des Bestimmens der inter-Frame regression ergibt.



WO 2020/109016 A1

Verfahren zum gemeinsamen Detektieren, Verfolgen und Klassifizieren von Objekten

Die Erfindung betrifft ein Verfahren zum gemeinsamen Detektieren, Verfolgen und Klassifizieren wenigstens eines Objektes in einer Videobildsequenz. Die Erfindung betrifft weiterhin ein Verfahren für ein maschinenbasiertes Training eines computerimplementierten Netzwerks zum gemeinsamen Detektieren, Verfolgen und Klassifizieren wenigstens eines Objektes in einer Videobildsequenz. Außerdem betrifft die Erfindung ein System zur Klassifizierung von Objekten, eine auf dem Medium abgespeicherte Computerprogrammprodukt sowie eine Anordnung.

Die automatisierte Identifizierung, Verfolgung und Klassifizierung von Objekten gewinnt zunehmend an Bedeutung in verschiedenen Bereichen wie beispielsweise bei der Überwachung von Personen, oder Orten, im Sport, bei der Bewegungsanalyse sowie im Bereich des autonomen Fahrens. Hierzu werden beispielsweise eines oder mehrere Objekte innerhalb eines Sichtfeldes einer Kamera identifiziert d.h. vom Hintergrund als getrennt wahrgenommen, gegebenenfalls in Klassen eingeteilt, deren Bewegungen über einen Zeitraum hinweg verfolgt und Aussagen darüber getroffen, wo sich die Objekte in Zukunft befinden werden. Im Bereich des autonomen Fahrens ist die Identifizierung und Klassifizierung von Objekten sowie deren Verfolgung über einen Zeitraum von wichtiger Bedeutung, um eine Risikoabschätzung vornehmen zu können oder als eine Entscheidungsgrundlage für darauf basierende Handlungen zu dienen. Für die Überwachung von Orten oder Personenströmen kann eine genaue Identifizierung von Personen anhand eindeutiger Merkmale notwendig sein, um diese auch nach einem Verlassen und späteren Wiedereintritt in den überwachenden Bereich sicher identifizieren zu können.

Für die automatisierte Detektion, Identifizierung und Verfolgung kommen häufig maschinenbasierte und trainierte Netzwerke zum Einsatz, die als „deep learning“ Netzwerke bezeichnet werden. Deep learning Netzwerke sind für komplexe Aufgabenstellungen besser geeignet als Netzwerke mit starren Regeln, da sich die große Anzahl an Variationen

nicht konkret in Regeln abbilden und programmieren lassen. Deep Learning Netzwerke bzw. Algorithmen werden für die Mustererkennung bereits häufig eingesetzt. Auch zum Detektieren und Identifizieren von Objekten in einem Sichtfeld werden darauf trainierte Netzwerke verwendet.

- 5 Dabei ist jedoch zu berücksichtigen, dass existierende Lösungen, vor allem bei niedrigen Bildwiederholungsraten Schwierigkeiten mit der Objektverfolgung bzw. der Wiederidentifizierung von Objekten haben können. Dies tritt vor allem dann auf wenn die Geschwindigkeit eines Objektes bezogen auf die Bildwiederholrate recht hoch ist, mit anderen
- 10 Worten sich zwischen den Einzelbildern stark ändert. Letzteres spielt unter anderem dann eine Rolle, wenn ein bereits identifiziertes Objekt den Bildbereich verlässt und anschließend wieder zurückkehrt oder wenn eine Verfolgung des Objektes aufgrund der niedrigen Bildrate fehlschlägt.
- 15 Somit wären weiterführende Lösungen, die eine höhere Flexibilität und bessere Genauigkeit bei der Identifizierung, Verfolgung und Klassifizierung von Objekten auch bei niedrigen Bildwiederholrate ermöglichen wünschenswert.

- In der hier vorgeschlagenen Lösung wird ein Algorithmus implementiert,
- 20 welche weniger speicherintensiv und schneller als konventionelle Lösungen ist. Dies wird erreicht, indem für jeden Zeitpunkt lediglich eine einzelne neuronale Netzwerkberechnung durchgeführt wird. Dazu werden die Detektion, eine Objektverfolgung sowie eine Merkmalsbestimmung und Klassifizierung eines jeden Objektes in einem Bild
- 25 gleichzeitig bestimmt. Insbesondere erfolgen auch die Bewegungsverfolgung eines Objektes sowie die Identifikation bzw. Re-Identifikation gleichzeitig.

- Eine derartige gleichzeitige Bestimmung erhöht die Performance gegenüber konventionellen Algorithmen, insbesondere bei niedrigen Bildwiederholraten bzw. großen Zeitabständen zwischen den Einzelbildern einer Videosequenz. Insbesondere erfolgt ein Training eines computer-
- 30 implementierten Deep-Learning Netzwerkes derart, dass das Detektieren, Verfolgen und Klassifizieren bzw. Identifizieren eines Objektes

oder mehrerer Objekte in Einzelbildern eine Videosequenz gleichzeitig erfolgt.

In einem Verfahren zum gemeinsamen Detektieren, Verfolgen und Klassifizieren wenigstens eines Objektes in einer Videobildersequenz aufweisend eine Vielzahl aufeinanderfolgende Einzelbilder mit einem trainierten computerimplementierten Netzwerk werden folgende Schritte ausgeführt:

Empfangen eines ersten Einzelbildes und eines darauffolgenden zweiten Einzelbildes;

- 10 Detektieren mindestens eines Objektes in dem ersten Einzelbild sowie mindestens eines Objektes in dem darauffolgenden zweiten Einzelbild;

Auswählen jeweils eines Objektes aus dem ersten und dem zweiten Einzelbild;

- 15 Ermitteln mindestens eines Klassifizierungssektors und einer Position für die ausgewählten Objekte aus dem ersten und dem zweiten Einzelbild;

Ermitteln eines Assoziierungswertes basierend auf dem ermittelten Klassifizierungsvektor und der Position; und

- 20 Erzeugen eines zeitlich konsistenten und eindeutigen Identifikationsvektors des mindestens einen Objektes für jedes Einzelbild in Antwort auf den ermittelten Assoziierungswertes. Hierbei ist für die Erzeugung des Assoziierungswertes eine von der Zeit zwischen dem ersten und zweiten Einzelbild abhängigen relative Gewichtung zwischen den ermittelten Klassifizierungsvektor und der Position vorgesehen. In diesem Zusammenhang wird unter dem Begriff Identifikationsvektor ein Vektor oder Wert verstanden, der eindeutig ist, so dass sich ein Objekt von einem anderen Objekt durch den Identifikationsvektor unterscheidbar ist.

- 30 Mit dem vorgeschlagenen Prinzip werden somit in zwei aufeinanderfolgenden Einzelbildern Objekte detektiert, identifiziert und über die Einzelbilder verfolgt. Dazu wird in vorteilhafter Weise die Position

sowie die Merkmals- bzw. Identifikationsvektoren der jeweiligen Objekte verwendet. Durch die Abhängigkeit des Assoziierungswertes von der Zeit und damit von der Bildwiederholrate zwischen dem ersten und dem zweiten Einzelbild wird eine relative Gewichtung zwischen dem Klassifizierungsvektor und den Positionsmerkmalen bzw. der Position des Objektes erreicht. In einem Aspekt enthält die Gewichtung eine Abhängigkeit invers von der Zeit zwischen den beiden Einzelbildern. Diese ist so gewählt, dass bei kleinen Bildraten verschiebt sich die relative Gewichtung hin zu dem Klassifizierungsvektor, sodass Objekte dennoch zuverlässig und sicher identifizierbar und verfolgbar sind. Mit anderen Worten wird das Gewichtungsfaktor für den Klassifizierungsvektor gegenüber der Gewichtung der Position umso größer je mehr Zeit zwischen den Einzelbildern verstreicht bzw. umso geringer die Bildwiederholrate wird.

Darüber hinaus wird durch den Klassifizierungsvektor und die Position ermöglicht, ein Objekt auch dann noch zu detektieren bzw. zu identifizieren, wenn es von anderen Objekten im darauffolgenden Frame teilweise verdeckt ist.

In diesem Zusammenhang kann zudem vorgesehen sein, eine Assoziierung eines Objektes in dem ersten bzw. zweiten Einzelbild auch mit Objekten in einem dritten Einzelbild vorzunehmen. Es kann Fälle geben, in denen eine Assoziierung eines Objektes in dem zweiten Teilbild mit einem Objekt des ersten Einzelbildes nicht möglich, das Objekt bleibt im zweiten Einzelbild somit „übrig“. In einem derartigen Fall, dass einem eindeutigen Identifikationsvektor eines Objektes eines zweiten Einzelbildes kein Vektor eines Objektes eines ersten Einzelbildes zuordenbar ist, kann der Identifikationsvektor des Objektes des zweiten Einzelbildes mit dem Identifikationsvektor eines Objektes eines dritten Einzelbildes verglichen werden, welches zeitlich vor dem ersten Einzelbild liegt.

In diesem Szenario wird durch den eindeutigen und zeitlich konsistenten Identifikationsvektor des Objektes eine Assoziierung mit Objekten auch in vorhergehenden Einzelbildern möglich. Es kann somit eine er-

neute Identifizierung bzw. eine Re-Identifizierung eines Objektes erfolgen, auch wenn dieses in einigen Einzelbildern nicht sichtbar ist. Ein typischer Anwendungsfall wäre beispielsweise durch ein Fahrzeug gegeben, welches von einem anderen Fahrzeug teilweise und über mehrere
5 Einzelbilder hinweg verdeckt ist und dann wiedererscheint. Durch das vorgeschlagene Prinzip wird das Fahrzeug re-identifiziert und nicht als neues Objekt erkannt.

In einem Aspekt der Erfindung umfasst der Schritt eines Detektierens zumindest eines Objektes ein Erzeugen einer bounding box, welches das
10 mindestens eine Objekt umgibt bzw. einschließt. Ebenso wird für die bounding box eine Vorhersage erzeugt, die sich auf die Veränderung der bounding box vom ersten zum zweiten Einzelbild hin bezieht. Weiterhin wird ein Geschwindigkeitsvektor für die bounding box des ersten Einzelbildes gebildet. Diese drei Schritte werden in einem Aspekt der
15 Erfindung vorzugsweise gleichzeitig durchgeführt. Ebenso kann für jedes Objekt eines Einzelbildes eine bounding box vorgesehen.

In einem weiteren Aspekt der Erfindung umfasst der Schritt des Auswählens wenigstens einen der folgenden Schritte, nämlich ein:

Auswählen der bounding box des ersten Einzelbildes und ein Auswählen
20 der bounding box des zweiten Einzelbildes;

Auswählen der Vorhersage sowie ein Auswählen der bounding box des zweiten Einzelbildes; und

Auswählen des Geschwindigkeitsvektors und ein Auswählen der bounding box des zweiten Einzelbildes.

25 Durch das Auswählen werden Kandidaten für den nachfolgenden Assoziierungsschritt ausgewählt. Sodann wird für jede der selektierten bounding boxen zwei Informationswerte extrahiert, nämlich die Position und ein dazugehöriger Klassifizierungsvektor. Gemäß einem Aspekt der Erfindung umfasst der Schritt des Ermitteln zumindest eines Klassi-
30 fizierungsvektors für das Objekt ein Erfassen von Merkmalen des Objektes, ein Berechnen eines eindeutigen Merkmalsvektors aus den erfassten Merkmalen sowie ein Klassifizieren des Objektes aus einer

Gruppe vorgegebener Klassen anhand der erfassten Merkmale oder anhand des berechneten Merkmalsvektor.

Hierbei kann die Gruppe vorgegebener Klassen unter anderem wenigstens eine der folgenden Klassen umfassen: Fußgänger, Autos, Busse, Fahrräder, Lastwagen, Traktoren, Tiere, Motorräder und Hindernisse, insbesondere stehende bzw. immobile Hindernisse.

In einem anderen Aspekt wird für den Schritt der Erzeugung einer zeitlich konsistenten und eindeutigen Identifikationsvektors ein spezielles Verfahren eingesetzt, beispielsweise ein „Hungarian combinatorial optimization“ Verfahren.

Das vorgeschlagene Verfahren teilt sich während der Durchführung des Verfahrens innerhalb des Basisnetzwerkes mehrere Faltungen für die unterschiedlichen Aufgaben, insbesondere für die Detektion der Objekte in den Einzelbildern, d. h. ein Erzeugen der bounding box das Erzeugen der Vorhersage für die bounding box und das Erstellen des Geschwindigkeitsvektors. Dadurch wird die Effizienz des Verfahrens erhöht und gleichzeitig die Hardwareanforderungen verringert. Insbesondere können in weiteren Schritten des Verfahrens die bounding boxen, welche das mindestens eine Objekt umgeben, für verschiedene weitere Aufgaben wie beispielsweise die Erzeugung der Vorhersage oder die Erzeugung des Klassifizierungsvektors sowie des Geschwindigkeitsvektors sowie zur Identifikation des Objektes verwendet werden. Im Gegensatz zu konventionellen Netzwerken, bei denen die verschiedenen Aufgaben durch unabhängige Netzwerke ausgeführt werden, wird nach dem vorgeschlagenen Prinzip jede Aufgabe gemeinsam bzw. parallel und eben nicht sequentiell ausgeführt und die Ergebnisse eines vorangegangenen Schrittes für die darauffolgenden Aufgaben gemeinsam bzw. parallel verwendet.

In einer Anordnung nach dem vorgeschlagenen Prinzip, vorzugsweise einer computerimplementierten Anordnung ist ein Zwischenspeicher für wenigstens ein erstes und ein zweites Einzelbild einer Videosequenz vorgesehen. Ein RPN-Modul („region proposal network“) ist mit dem Zwischenspeicher gekoppelt. Das RPN-Modul ist ausgeführt zur Erzeugung einer ersten Vielzahl von Objektvorschlägen aus dem ersten Einzelbild

und einer zweiten Vielzahl von Objektvorschlägen aus dem zweiten Einzelbild, welche im Folgenden als Roi's oder „region of interest“ bezeichnet werden können. Mit dem RPN-Modul ist ein Bestimmungs- und Identifikationsmodul verbunden. Dieses Modul ist ausgestaltet aus der ersten Vielzahl und der zweiten Vielzahl einen Identifikationsvektor für jedes Objekt innerhalb der ersten und zweiten Vielzahl zu erzeugen. Daneben ist das Modul ausgeführt parallel zu dem vorstehenden, eine Klassifizierung der jeweiligen Objekte vorzunehmen sowie eine klassenspezifische bounding box regression für jeden Vorschlag zu erzeugen.

Ebenfalls mit dem RPN-Modul ist ein Vorhersagemodul gekoppelt und somit hinsichtlich der Funktionalität parallel zu dem Bestimmungs- und Identifikationsmodul angeordnet. Dieses ist zu einer Korrelationsanalyse für ein Objekt in einem in dem ersten Einzelbild bezüglich dessen Position mit entsprechenden Objekten eines zweiten Einzelbildes ausgeführt. Mit anderen Worten ist das Vorhersagemodul ausgeführt, eine Vorhersage über eine Entwicklung eines Objektvorschlages vom ersten Einzelbild in Hinsicht auf das zweite Einzelbild zu treffen.

Dazu ist das Vorhersagemodul in einem Aspekt der Erfindung mit einer Korrelationsschicht ausgeführt. Diese ist ausgebildet, eine Korrelation zwischen einem Objektvorschlag in einem ersten Einzelbild und benachbarten Fenstern in einem zweiten Einzelbild zu bestimmen.

Ein anderer Aspekt dieser Erfindung beschäftigt sich mit dem Verfahren für ein maschinenbasiertes Training eines computerimplementierten Netzwerkes zum gemeinsamen Detektieren, Verfolgen und Klassifizieren wenigstens eines Objektes in einer Videobildsequenz. Die Videobildsequenz verweist hierzu eine Vielzahl aufeinanderfolgende Einzelbilder auf.

Das Verfahren umfasst in einem ersten Schritt ein Bereitstellen einer Vielzahl von Einzelbildpaaren, wobei jedes Einzelbildpaar ein erstes und ein darauffolgendes zweites Einzelbild umfasst und jedes Einzelbild wenigstens einen Objektvorschlag auch „regions of interest“ oder Roi genannt aufweist, den je ein Klassenidentifikationsvektor eindeutig zugeordnet ist. In einer zweckmäßigen Ausgestaltung können der

wenigstens eine Objektvorschlag jeweils ein Objekt aufweisen, das durch den Klassenidentifikationsvektor eindeutig in seiner Klasse identifizierbar ist.

Für jeden Objektvorschlag wird nun ein Merkmalsvektor bestimmt, welchen der Klassenidentifikationsvektor zugeordnet ist. Der ermittelte Merkmalsvektor kann so einer Klasse zugeordnet werden. Ebenso wird ein Identifikationsvektor sowie eine spezifische bounding box Regression bestimmt. Letztere ist spezifisch für den Klassenidentifikationsfaktor. Anschließend wird für korrelierte Merkmalskarte zwischen dem ersten Einzelbild und dem darauffolgenden zweiten Einzelbild erzeugt sowie für jeden Objektvorschlag eine Interframe Regression bestimmt.

Für das Training des computerimplementierten Netzwerks wird nun ein kombinierter Fehler minimiert, wobei dieser sich aus den Fehlern beim Bestimmen des Klassenidentifikationsvektors, beim Bestimmen des wenigstens einen Identifikationsvektors und beim Bestimmen der spezifischen bounding box Regression sowie der Interframe Regression ergibt.

Durch das Training des Netzwerks mit einem kombinierten Fehler der oben genannten Art wird eine Detektion eines Objektes, eine Verfolgung sowie eine Merkmalsidentifikation für das Objekt gleichzeitig trainiert. Da der Klassenidentifikationsvektor für den Objektvorschlag und damit vorzugsweise für das Objekt innerhalb des Objektvorschlags fest vorgegeben ist, bedingt das Verfahren ein Trainieren von Merkmalen, die der jeweiligen Klasse entsprechen.

In einem weiterführenden Aspekt umfasst der Schritt für jeden des wenigstens einen Objektvorschlags ein Zusammenlegen von Merkmalen einer vordefinierten Größe aus den Einzelbildern mittels eines ROI-Align Verfahrens. Dieses erfahren kann ein R-CNN Verfahren enthalten oder Teil davon sein. Die vordefinierte Größe kann dabei wenige Pixel beispielsweise 7x7 Pixel betragen. Die zusammengelegten Merkmale werden mittels eines Faltungsblocks gefaltet und nachfolgend zu einem globalen Vektor zusammengefasst. Zum Bestimmen des Identifikationsvektors kann ein online instant Matching Loss (OIM) Verfahren verwendet

werden. Insbesondere wird mit diesem Verfahren die korrekte Bestimmung des Merkmalsvektors und/oder des Identifikationsvektors gelernt. In einem weiteren Schritt umfasst das Bestimmen des Identifikationsvektors das Anwenden einer voll verbundenen Schicht einer definierten Größe um einen gegenüber dem globalen Vektor niedrigeren Identifikationsvektor zu erhalten.

In einer weiteren Ausgestaltung wird ein Klassifikationsscore ermittelt, welcher für die korrekte Klassifizierung des Objektes in dem Objektvorschlag benötigt wird. Dieser kann bestimmt werden durch Anwenden zweier vollverbundener Schichten der Größe $(C + 1)$ und $x(C+1)$, wobei $(C + 1)$ die Anzahl der Klassenidentifikationsvektoren und damit der Klassen einschließlich des Bildhintergrundes ist.

Um den kombinierten Fehler aus den einzelnen vorhergehenden Bestimmungen zu minimieren, können verschiedene Verfahren angewendet werden. Beispielsweise kann ein Software Max Cross Entropie loss Verfahren für die Bestimmung des Fehlers bei der Klassifizierung angewendet werden. Ein Smooth L1-Loss Verfahren dient zur Bestimmung der bounding box Regression Fehlers sowie zur Bestimmung des Fehlers in einer Inter-frame Regression.

Ein weiterer Aspekt betrifft den Schritt des Bestimmens der korrelierten Merkmalskarte. Dazu kann eine Korrelation zwischen der ersten Merkmalskarte des ersten Einzelbildes und benachbarten Ausschnitten in einer zweiten Merkmalskarte des zweiten Einzelbildes ermittelt werden. Mit anderen Worten wird eine Korrelation zwischen Merkmalen der ersten und der zweiten Merkmalskarte dadurch erfasst, dass Merkmalen an einer Position der ersten Merkmalskarte des ersten Einzelbildes mit den Merkmalen einer zu der ersten Position benachbarten Position verglichen werden.

Ein weiterer Aspekt betrifft eine Anordnung, vorzugsweise eine computerimplementierte Anordnung. Diese umfasst einen Zwischenspeicher für wenigstens ein erstes und ein zweites Einzelbild einer Videosequenz; sowie ein RPN-Modul zur Erzeugung einer ersten Vielzahl unabhängiger Objektvorschläge aus dem ersten Einzelbild und einer zweiten Vielzahl unabhängiger Objektvorschläge aus dem zweiten Einzelbild.

Die Anordnung enthält ein mit einem Vorhersagemodul gemeinsam trainiertes Bestimmungs- und Identifikationsmodul sowie das Vorhersagemodul, das mit dem Bestimmungs- und Identifikationsmodul gemeinsam trainiert ist. Alle Module sind mit dem RPN-Modul gekoppelt.

- 5 Das trainierte Bestimmungs- und Identifikationsmodul ist ausgeführt ist, Objekte aus der ersten Vielzahl und der zweiten Vielzahl von Objektvorschlägen zu identifizieren und im ersten und zweiten Einzelbild zu detektieren und welches weiter ausgeführt, aus den Objektvorschlägen eine Vorhersage der Position eines Objektes im ersten Einzelbild und im zweiten Einzelbild zu treffen. Das Vorhersagemodul, ist ausgeführt, eine inter-Frame Vorhersage für jedes Objekt im ersten Einzelbild basierend auf der ersten Vielzahl von Objektvorschlägen und der zweiten Vielzahl von Objektvorschlägen zu treffen.

- 10 Durch das gemeinsame Training und die End-to-End Optimierung wird die Effizienz und die Erkennungs-, Klassifizierung und Verfolgungsrate gerade bei kleinen bildwiederholraten signifikant verbessert.

- 15 In einem weiterführenden Aspekt ist dem RPN-Modul eine BaseNet Architektur vorgeschaltet. Diese ist ausgeführt, eine Vielzahl von Basismerkmalen für das erste und das zweite Einzelbild zu ermitteln, welche dem RPN-Modul zur Erzeugung der Objektvorschläge zuführbar sind.

- 20 In einem anderen Gesichtspunkt ist das trainierte Bestimmungs- und Identifikationsmodul ausgeführt, für jeden Vorschlag Merkmale einer vorbestimmten Größe aus den Basismerkmalen zusammenzufassen, um einen mehrdimensionalen Merkmalsvektor bereitzustellen. Zudem kann es ausgeführt sein mit Hilfe des Merkmalsvektors eine Klassifizierungsparameter des Objektes im Objektvorschlag zu bestimmen. Hierbei kann Klassifizierungsparameter aus einer Menge vorbestimmter Klassen ausgewählt ist. Dadurch lässt sich die Anordnung auch auf ganz spezielle Probleme und Situationen trainieren und optimieren.

- 25 In einem weiteren Aspekt ist das trainierte Bestimmungs- und Identifikationsmodul ausgeführt, ein Objekt in dem Objektvorschlag einen eindeutigen Identifikationsvektor zuzuordnen. Dies kann dann zweckmäßig sein, wenn das Objekt über einige wenige Einzelbilder hinweg

verschwindet und dann wiederauftaucht. Durch den Identifikationsvektor lässt sich eine Re-Identifizierung vornehmen, da jedes Objekt einen eindeutigen Identifikationsvektor enthält. Ebenso kann das trainierte Bestimmungs- und Identifikationsmodul ausgeführt sein, für
5 jeden Objektvorschlag eine bounding box regression zu erzeugen.

Ein anderer Aspekt betrifft das trainierte Vorhersagemodul. Dieses kann ausgeführt sein, anhand der Basismerkmale des ersten und zweiten Einzelbildes die Regressionsziele der detektierten Objekte vom ersten zum zweiten Einzelbild vorherzusagen.

- 10 Im Folgenden wird die Erfindung anhand mehrere Ausführungsbeispiele und unter Bezugnahme auf die Zeichnungen im Detail erläutert. So zeigen:

Figur 1 ein Flussdiagramm des vorgeschlagenen Systems;

- 15 Figur 2 ein Computersystem zur Durchführung des vorgeschlagenen Verfahrens;

Figur 3 eine Videosequenz mit Einzelbildern zur Darstellung von Ergebnissen mit einem Netzwerk nach dem vorgeschlagenen Prinzip;

Figur 4 eine Ausführungsform einer computerimplementierten Anordnung mit den verschiedenen funktionalen Modulen;

- 20 Figur 5 eine schematische Darstellung des Modules PredNet zur Bestimmung einer Vorhersage der Bewegung oder Position eines Objektes;

- Figur 6 ein Ausführungsbeispiel des Verfahrens zum gemeinsamen Detektieren, Verfolgen und Klassifizieren wenigstens eines Objektes
25 in einer Videobildsequenz;

Figur 7 eine Ausführungsform eines Verfahrens zum Trainieren eines computerbasierten Netzwerks nach dem vorgeschlagenen Prinzip.

- Eine automatisierte Objektdetektion und Erkennung gerade im Bereich von Videoüberwachung und Robotertechnologie zeigt seit langem verschiedene Herausforderungen. Von Bedeutung ist auch die Verfolgung
30 eines Objektes, die oftmals notwendig wird, um mit dynamischen Szenen

umgehen zu können. Eine Personenidentifikation und das Verfolgen einer an sich identifizierten Person oder Objektes über mehrere Sichtfelder hinweg hat in den letzten Jahren an Bedeutung gewonnen.

Figur 1 zeigt in diesem Zusammenhang einen einfachen Flow Chart für ein erfindungsgemäßes System, bei dem verschiedene Aufgaben gemeinsam und gleichzeitig ausgeführt werden. Eine Kamera 1 nimmt hierbei eine Vielzahl von Bildern auf, beispielsweise in Form einer Videosequenz. Die Bilder werden einem Netzwerk 2 zugeführt, welches im Folgenden näher erläutert wird. Dieses detektiert und identifiziert verschiedene Objekte, die von der Kamera in ihrem Gesichtsfeld aufgenommen werden und übergibt diese an ein Verfolgungsmodul 3. Das Verfolgungsmodul erzeugt aus den Informationen des Netzwerkes 2 für jedes der Objekte Trajektorien über die Zeit, die anschließend an verschiedene Anwendungen 4 zur Auswertung übergeben werden können.

Netzwerk 2 umfasst dabei mehrere funktionale Elemente, die mit den Bezugszeichen 5, 6 und 7 versehen sind. Zu den funktionalen Elementen gehört unter anderem ein Modul oder eine Funktion zur Identifikation eines Objektes, ein Modul oder Funktion zu deren Detektion innerhalb des übergebenen Bildes sowie ein Modul oder Funktion für eine Bewegungs- oder Positionsvorhersage, um das Objekt über einen längeren Zeitraum und mehrere Bilder hinweg verfolgen zu können.

Eine Detektion eines Objektes innerhalb eines Bildes erfolgte mit dem Modul 7. Detektoren dieser Art arbeiten entweder „region proposal“ Detektoren oder als Einzeldetektoren. Bei region proposal basierten Detektoren wird das Einzelbild in verschiedene Regionen bzw. Bereiche aufgeteilt und diese entsprechend mit verschiedenen Verfahren verarbeitet. Dazu ist ein sogenanntes RPN-Netzwerk vorgesehen, welches einen Satz von Objektvorschlägen (diese sind häufig mit den Regionen gleichgesetzt) erzeugt, die an ein neurales Netzwerk zur Klassifikation und Regression weitergegeben werden. Zur Erzeugung von Objektvorschlägen werden „fast R-CNN“, „faster, R-CNN“ und „RFCN“-Verfahren oder generell auch ROI-Align Verfahren benutzt. Andere Varianten von Objektdetektoren erzeugen keine Objektvorschläge, sondern bestimmen

„bounding boxes“ um die Objekte und die entsprechenden Klassifizierungen des Bildes für die einzelnen Objekte direkt. Derartige Objektdetektoren basieren beispielsweise auf YOLO Verfahren.

Für Videosequenzen, d. h. eine Vielzahl aufeinanderfolgende Bilder können eine Erfolgsrate für eine Objektdetektion signifikant verbessert werden, wenn die entsprechenden Objekte beispielsweise über mehrere Einzelbilder hinweg sichtbar sind. Demgegenüber erschwert eine Bewegung von Objekten in einer Videosequenz von einem Bild auf das nächste die Detektion, da sich die Position des Objektes durch die Bewegung ändert. Zur Verringerung dieses Problems wird in herkömmlichen Detektoren und Verfahren eine Detektion des Objektes zusammen mit einer Verfolgung desselben durchgeführt, der Fehler bestimmt und minimiert. Derartige Ansätze können nicht nur für einzelne Objekte, sondern auch für mehrere Objekte innerhalb eines Bildes verwendet werden.

Neben einer Objektdetektion steht aber auch eine Identifikation und Einteilung in verschiedenen Klassen bzw. eine Re-identifikation im Vordergrund. Diese Forderung ergibt sich aus verschiedenen Anwendungen, bei denen es notwendig ist, das diktierte Objekt auch hinsichtlich seiner Eigenschaften besser zu charakterisieren. Beispielsweise wäre es im Straßenverkehr erforderlich, einen Bus bzw. ein Auto von einem Fußgänger oder einem Fahrradfahrer ausreichend genau unterscheiden zu können. Nur dann kann eine Risikoabschätzung ausreichend gut vorgenommen und die korrekten Handlungen hierauf selektiert werden.

In anderen Anwendungsfällen ist eine Re-identifikation eines Objektes erforderlich. Beispielsweise könnte eine durch mehrere Gesichtsfelder von Kameras wandernde Person über diese Felder hinweg nicht nur detektiert, sondern auch identifiziert und verfolgt werden. Gleiches gilt auch im Straßenverkehr, wenn beispielsweise ein Objekt während einer kurzen Zeit durch ein anderes Objekt verdeckt wird und es nach dem Wiederauftauchen als bereits bekanntes Objekt deklariert werden soll.

Zur Identifikation bzw. Re-identifikation und anschließender Klassifizierung wird vorgeschlagen, einen eindeutigen Merkmalsvektor für jede Identität und damit Klasse zu trainieren. Dieser Merkmalsvektor sollte vorzugsweise invariant gegenüber Veränderungen der Kameraposition, der Beleuchtung, des Hintergrundes, der Farben und der perspektivischen Ansicht sein.

Verschiedene hierzu vorgeschlagene Verfahren benutzen geschnittene bounding boxen mit dem zu charakterisierenden Objekt. Es wurde festgestellt, dass eine Wiederidentifizierung und Detektion von Objekten in ganzen Bildern möglich ist. Mit einem eindeutigen Identitätsvektor kann zusätzlich eine Verfolgung des Objektes über mehrere Einzelbilder hinweg verbessert werden.

Schließlich umfasst das Netzwerk 2 neben einer Funktion zur Detektion 7 und Identifizierung 6 ein Vorhersagemodul 5, welches ermöglicht, eine Vorhersage über die Bewegung oder die Position eines Objektes in einem darauffolgenden Einzelbild zu treffen. Eine derartige Vorhersage kann nicht nur dazu verwendet werden, die Detektion bzw. Verfolgung eines Objektes zu verbessern, sondern auch für eine Risikoanalyse hinsichtlich der Bewegungsrichtung eines Objektes, beispielsweise im Bereich des autonomen Fahrens oder der Robotertechnologie verwendet werden.

Ergebnisse bzw. Resultate, die mit einem derartigen Netzwerk gewonnen werden, sind in Figur 3 dargestellt. Diese zeigen drei aufeinanderfolgende Einzelbilder eines Verkehrs auf einer Straße zu den Zeiten $t-20$, $t-10$ und t . Im ersten Einzelbild zum Zeitpunkt $t-20$ hat das Netzwerk insgesamt 4 verschiedene Objekte erkannt. Dazu gehört ein Bus 32, zwei auf der ganz linken Spur hintereinanderfahrende PKW 33 und 34, sowie ein hinter den Bus auf die Straße einbiegender kleiner Transporter 31. Jedes der identifizierten Objekte ist zudem mit einer entsprechenden bounding box, d.h. einem Rechteck umgeben, welches die äußeren Abmessungen des jeweiligen Objektes, bzw. Fahrzeugs umgibt. Wie im Bild angedeutet, sind die bounding boxen zudem mit leicht unterschiedlichen Farben dargestellt, was einer eindeutigen Charakterisierung und Klassifizierung entspricht. Im darauffolgenden Bild

haben sich die Fahrzeuge fortbewegt. Durch die Identifizierung und das Folgemodul des Netzwerks ist das System in der Lage, die Objekte weiter zu verfolgen und gleichzeitig als Objekte zu identifizieren, die bereits im vorangegangenen Einzelbild vorhanden waren. Im Besonderen, sind die beiden PKWs 33 und 34 etwas näher gekommen, das Netzwerk erkennt und ordnet Ihnen weiterhin die gleiche ID (Identifikation) und Klasse zu. Zusätzlich hat das Netzwerk ein neues Objekt 35 erkannt, eine eindeutige Identifikation diesem zugeordnet und es als weiteren PKW klassifiziert.

Demgegenüber ist im Bild t-10 der im vorherigen Bild erkannte Transporter hinter dem Bus verschwunden und wird von diesem Objekt somit verborgen. Entsprechend hat das Netzwerk 2 das Objekt verloren, d.h. eine Verfolgung des Transporters ist fehlgeschlagen. Im letzten Einzelbild t ist Fahrzeug 33 mittlerweile aus dem Gesichtsfeld verschwunden. Gleichzeitig ist der Transporter 31 durch den Bus nicht mehr verdeckt, sondern vom Netzwerk erneut erkannt worden. Gemäß dem vorgeschlagenen Prinzip ist das Netzwerk 2 nicht nur in der Lage, das Objekt erneut zu detektieren und richtig zu klassifizieren, sondern es auch als bereits bekannte Fahrzeug zu identifizieren. Dies ist möglich, da ein Training des Netzwerks gemäß dem vorgeschlagenen Prinzip gleichzeitig erfolgt.

Fig. 4 zeigt ein Model nach dem vorgeschlagenen Prinzip. Eine Objekterkennung und -verfolgung wird durch ein gemeinsames Training der drei Aufgaben Erkennung, Vorhersage und Neu- bzw. Re-Identifizierung adressiert. Unter dem Begriff „gemeinsam“ wird hierbei ein paralleles oder kombiniertes Training verstanden. Es ist also nicht sequentiell, d.h. es werden die einzelnen Aufgaben nicht unabhängig voneinander trainiert, sondern in einer Kombination. Im Bereich des maschinenbasierten Lernens bedeutet dies, dass der Fehler beim Trainieren für jede Aufgabe nicht einzeln und separat minimiert wird, sondern eine Kombination aus allen auftretenden Fehlern.

Im Gegensatz zu traditionellen sequentiellen Ansätzen, schlägt die Erfindung vor bei der Erkennung, Detektion, Identifizierung, Klassifizierung und Verfolgung eine einheitliche End-to-End-Architektur

vor, die diese verschiedenen Aufgaben gemeinsam und zusammen optimiert.

Die vorgeschlagene Modellarchitektur ist in Figur 4 dargestellt und enthält die zwei Ausgangszweige DetIdenNet und PredNet. Letzteres Modul wird hinsichtlich seines Aufbaus auch in Figur 5 gezeigt. Hierbei gibt der erste Zweig oder Modul DetIdenNet einen Identitätsvektor ID, eine box regression BR und einen Klassifizierungswert CS aus. Die Architektur benötigt zwei aufeinander folgende Frames I^{t-1} und I^t zu den Zeitpunkten $t-1$ und t . jedoch kann statt $t-1$ (d.h. das vorherige Bild) auch jeder andere Zeitpunkt $t-\delta$ verwendet werden. Diese entsprechen zwei aufeinanderfolgende Einzelbilder. Die beiden Einzelbilder, auch Frames genannt werden zuerst durch das Modul BaseNet geleitet, welches die ersten vier Blöcken (Conv1 - 4) einer ResNet-Architektur umfasst. Dabei präsentieren f^{t-1} und f^t ($\in h \times w \times d$) die Basismerkmale für die beiden Einzelbilder. h , w und d sind Höhe, Breite und Anzahl der Kanäle der Basismerkmale. Zusätzlich zu diesen Funktionen wird dem BaseNet Module ein Regionsvorschlagsnetzwerk (RPN) oder RPN-Modul nachgeschaltet, um unabhängige Objektvorschläge von beiden Einzelbilder zu erhalten, die an die beiden Module DetIdenNet und PredNet weitergeleitet werden.

Für jeden Vorschlag aus dem RPN-Modul bündelt das Modul DetIdenNet Merkmale der Größe 7×7 aus den jeweiligen Basismerkmalen zusammen, indem es ein ROI-Align Verfahren verwendet. Die zusammengefassten Merkmale werden dann durch eine Reihe von Faltungen aus dem letzten Faltung Block (Conv5) von ResNet geleitet gefolgt von einer globalen Durchschnittsbildung. Dadurch entsteht ein 2048-dimensionales Merkmal für jeden Vorschlag. Dieses Merkmal wird dann an zwei Geschwisterzweige weitergereicht zur Identifikation bzw. Detektion.

Der Identifizierungszweig wendet zuerst eine vollständig verbundene Schicht der Größe 256 an, wodurch ein Identitätsmerkmal mit geringerer Dimension entsteht. Der Detektionsszweig wendet zwei parallele, vollständig verbundene Schichten an der Größe $(C + 1)$ und $4 \times (C + 1)$. Dies ergibt einen Klassifizierungswert sowie eine klassenspezifische bounding box Regression für jeden Vorschlag. Dabei ist $(C + 1)$ die

Anzahl der Klassen einschließlich des Hintergrunds. Für das Training wird für den Detektionszweig ein „Softmax Cross Entropie Loss“ Verfahren zur Klassifizierung und ein Smooth-L1-Loss Verfahren für die Bounding box Regression verwendet. Um die Einbettung der 256-dimensionalen Identifizierungsfunktion zu erlernen und zu trainieren wird
 5 ein Online Instance Matching Loss (OIM) Verfahren benutzt.

Angesichts der Basismerkmale f^{t-1} und f^t der Größe $h \times w \times d$ der beiden Einzelbilder, zielt PredNet darauf ab, die Regression für die Objekterkennung vorherzusagen und zwar vom ersten Bild zum zweiten. Um
 10 dies zu erreichen, verwendet PredNet zunächst eine Korrelationsschicht, welche die Korrelation jedes Merkmals $f^{t-1}(x, y)$ in der ersten Merkmalskarte mit ihrem benachbarten $(2n + 1) \times (2n + 1)$ Fenster in der zweiten Merkmalskarte verwendet. Dieser Vorgang produziert eine Merkmalskarte C der Größe $h(2n + 1) \times w(2n + 1)$ wie in der folgenden
 15 Gleichung gezeigt:

$$C(x, y) = \sum_d f^{t-1}(x, a, d) 1(n, n) \odot N[f^t(x, y, d)]$$

Hierbei sind $f^{t-1}(x, y, d)$ und $f^t(x, y, d)$ Skalarwerte der räumlichen Position x, y und des Kanals d in der Merkmalskarte f^{t-1} und f^t . $1(n, n)$ ist eine $(2n + 1) \times (2n + 1)$ Matrix von Einsen, die verwendet wird,
 20 um den Skalarwert $f^{t-1}(x, y, d)$ zu einer $(2n + 1) \times (2n + 1)$ Matrix zu wiederholen. $N[f^t(x, y, d)]$ ist die $(2n + 1) \times (2n + 1)$ Nachbarschaftsmatrix von $f^t(x, y, d)$. Die obige Gleichung berechnet zuerst die elementweise Multiplikation der Matrix $f^{t-1}(x, y, d) 1(n; n)$ mit der
 25 Nachbarschaftsmatrix $N[f^t(x, y, d)]$ und summiert sie dann entlang der Kanalabmessung. Es sei darauf hingewiesen, dass $C(x, y)$ ein Block der Größe $(2n + 1) \times (2n + 1)$ ist und die Korrelation des Merkmals $f^{t-1}(x, y)$ mit der $(2n + 1) \times (2n + 1)$ Nachbarschaft der Matrix in $f^t(x; y)$ angibt. Die korrelierte Merkmalskarte C wird dann an eine Patch-Zusammenfassungsschicht übergeben, die jedes $(2n + 1) \times (2n + 1)$ -Fenster unter
 30 Verwendung einer Filterfaltung der Größe $(2n + 1) \times (2n + 1)$ mit Schrittweite $2n + 1$ zusammenfasst und 512 Ausgangskanäle ausgibt. Diese zusammengefasste Merkmalskarte der Größe $(h \times w \times 512)$ wird dann mit

den ursprünglichen Basismerkmalen f^{t-1} und f^t verkettet, und eine 1×1 Faltung durchgeführt, um d Ausgangskanäle zu erhalten. Dies erlaubt es, den standard Conv5 Block der ResNet-Architektur zu benutzen. Diese Merkmale werden dann mit einem ROI-Align Verfahren weiter verarbeitet und zusammen mit den Detektions-bounding boxen des ersten Einzelbildes, gefolgt von einem Conv5-Block von ResNet und eine vollständig verbundene Schicht um eine Regression $\Delta^t = (\Delta_x^t, \Delta_y^t, \Delta_w^t, \Delta_h^t)$ für jeden Verfolgungs- oder Track-Objektvorschlag. Während des Training wird für PredNet ein Smooth L1 loss Verfahren zwischen den Grundwahrheitszielen und den vorhergesagte Ziele durchgeführt.

Um die Zweige DetIdenNet und PredNet der Anordnung Netzwerks gemeinsam zu trainieren, wird ein kombinatorischer Gesamtfehler eingesetzt, der sich aus dem Klassifikationsverlust (L_{cls}), dem Regressionsverlust (L_{reg}), einem Regressionsverlust für die Zwischenbilder bounding boxen (L_{tra}) und Verlust bei der Identifizierung (L_{iden}) zusammensetzt. Der Gesamtverlust für eine Anzahl von N Objektvorschläge (Rois) ergibt sich somit zu:

$$L = \frac{1}{N} \sum_{j=1}^N L_{cls}(p_{j,c^*}) + \lambda_1 \frac{1}{N_{fg}} \sum_{j=1}^N [c_j^* > 0] L_{reg}(b_j, b_j^*) + \lambda_2 \frac{1}{N_{tra}} \sum_{j=1}^{N_{tra}} L_{tra}(\Delta_j^t, \Delta_j^{*,t}) + \lambda_3 \frac{1}{N_{iden}} \sum_{j=1}^{N_{iden}} L_{iden}(q_{j,i^*})$$

Hierbei ist für jeden Objektvorschlag j , $p_{j,c}$ die vorhergesagte Klassifizierungswahrscheinlichkeit seiner Grundwahrheitsklasse c^* und b_j , b_j^* sind die vorhergesagten und Grundwahrheits-Bounding-Box-Ziele. In dem Tracking-Verlust stehen die Terme Δ_j^t und $\Delta_j^{*,t}$ für die normalisierten Interframe, vorhergesagten und Grundwahrheits-Bounding-Box-Ziele.

Für das Training wurden lediglich N_{tra} Grundwahrheiten (Track Rois) vom ersten Einzelbild $t-1$ verwendet, die auch entsprechende Kästchen (gleiche Id) im Einzelbild t haben. Bei Interferenz werden alle Detektionen (N) von der ersten Einzelbild Verfolgungsobjektvorschläge (track Rois) verwendet. Im Identitätsverlust $q_{j,i}$ ist die vorhergesagte

Wahrscheinlichkeit des Objektvorschlags j für seine Grundwahrheitskennung i . Dieser Verlust ist nur für die Objektvorschläge im Vordergrund berechnet, denen auch eine Grundwahrheitsidentifizierung i^* (N_{iden}) zugeordnet ist. Die Verlustgewichtung λ_x sind auf 1 gesetzt.

5 Für das Verfolgen wird ein einfacher Tracking-by-Detection-Algorithmus implementiert, der die Erkennungshypothesen mit Zielbahnen durch Optimierung des Hungarian association Verfahrens verknüpft. Hierzu wird die Cosinus-Ähnlichkeit der ID-Merkmale und / oder der IOU Überlappung berücksichtigt, um die Affinitätsmatrix für den Hungarian
10 association zu berechnen. Zur Vorhersage der Ziel-Begrenzungs-Boxen im darauffolgenden Einzelbild wird ein einfaches lineares Bewegungsmodell und bounding box Vorhersagen aus dem PredNet Modul benutzt. Ziel ist es, jede erhaltene Detektion einer korrekten Trajektorie aus dem verfügbaren Satz von Trajektorien zum Zeitpunkt $t-1$ zuzuordnen.

15

Das Verfahren benutzt hierzu einen Tracking buffer. Dieser enthält die Trajektorieninformation zu allen Trajektorien zum $t-j$. Die Trajektorieninformation umfassen (a) die bounding box am Trajektorienkopf, den Merkmalsvektor am Trajektorienkopf zu den Durchschnittsgeschwindigkeitsvektor des Ziels.
20

Um eine korrekte Trajektorie einer Detektion zuzuordnen, wird eine zweiteilige Grafik zwischen alle Detektionen im aktuellen Bild t und alle Trajektorien im Bild $t-1$ erstellt, gefolgt von einem dem hungarian association Verfahren, um eine optimale Menge der Eins-zu-Eins-Zuordnungen zu finden. Die Assoziierungsgewichtung der Graphenkanten werden durch eine Affinitätsmatrix angegeben.
25

Dazu lassen sich zwei Arten von Assoziationsmetriken unterscheiden. Zum einen ein Verbund von bounding-box-intersection-union (IoU), der im Wesentlichen auf der Überschneidung von bounding boxen beruht und
30 zum anderen eine Assoziierung basierend auf einer ID-Merkmalseinbettung. Die Cosinus Ähnlichkeit wird verwendet, um die Assoziation zu berechnen. Im Vergleich zu ID-Funktionen ermöglicht die IoU-Metrik eine ergänzende räumliche Vorverfolgung für den Verfolgungsschritt.

Damit wird vermieden, falsche Objekte abzugleichen, die ähnliches Aussehen haben, zum Beispiel Autos des gleichen Modells / der gleichen Marke oder Personen, die ähnliche Kleidung haben. In diesem Zusammenhang wurde gefunden, dass eine einfache Kombination der beiden Metri-

5 riken Verbesserungen gegenüber einer getrennten Benutzung der Metriken.

Das folgende Bild zeigt den Ablauf des Verfahrens zur Nachverfolgung von Objekten:

Algorithm 1 Tracking algorithm

Input : Frame at time t , Detections D^t , ID-features f^t , predictions $Pred^t$, $buffer\_size$, w_1 , w_2

Initialization: $j = 1$

1. Get trajectory information T^{t-j} at $t - j$ from the trajectory buffer.
2. Calculate the overall affinity matrix, A^t (See Eq. 3).
3. Use hungarian method to find optimal assignment between T^{t-j} and D^t .
4. Update trajectories T^t at frame t for assigned detections, $D_a^t \in D^t$
5. If $j < buffer\_size$ do:
 $j = j + 1$
 for unassigned detections, $D_u^t \in D^t$, repeat steps 1 to 5 with $w_1 = 0$ and $w_2 = 1$.
6. Assign new ID to all D_u^t , and initialize new trajectories.
7. For all unmatched trajectories, T_u^{t-1} :
 add prediction, $Pred_u^t$ if available from network
 else propagate the trajectory based on a linear motion model.

10 Um das Verfolgungsmodul robuster gegenüber Verdeckungen (Okklusionen) und fehlenden Detektionen oder vorhersagen zu machen, wird ein Puffer verwendet, in dem Trajektorien gehalten werden können, die mit keiner neuen Detektion verbunden waren. Für die Detektionen bei einem Einzelbild t die keiner Trajektorie des Bildes $t-1$ zugeordnet werden

15 können, wird ihr Identifikationsvektor ID für die nicht zugewiesenen bzw. in dem Puffer gehaltenen Trajektorien verglichen. Dies sind

Trajektorien von vorangegangenen Bilder T-2, t-3 usw. Im Puffer gehaltene Trajektorien des Einzelbildes t-2 sind solche, bei denen es keine Assoziierung mit einer Detektion im Bild t-1 gab. Daher sind solche zeitlich unterbrochenen Trajektorien verfügbar für den Abgleich für Bild t.

Der Puffer für diese gehaltenen Trajektorien hat eine vordefinierte Größe die somit in einem Zusammenhang steht, wie weit in der Zeit zurück Trajektorien mit neu detektierten Objekten assoziiert werden können. Um die Fragmentierung in den Trajektorien zu reduzieren, lässt sich ein einfaches Model verwenden, bei dem die pausierenden Trajektorien für einen kurzen Zeitraum verbreitert werden zusätzlich zu den Vorhersagen aus dem PredNet Modul.

Figur 2 zeigt eine Computeranordnung mit Kamerasystem, auf dem das vorgeschlagenen Verfahren ablaufen und auch Anordnung mit den verschiedenen funktionalen Modulen implementiert werden kann. Insbesondere ist das Computersystem für die Implementierung von neuronalen Netzwerken ausgebildet, die durch maschinenbasiertes Lernen implementiert sind. Zu diesem Zweck enthält das Computersystem 20 ein oder mehrere Prozessoren 21 und 22. Die Prozessoren sind zu massiven parallelen Verarbeitung von Daten ausgeführt. Dazu eignen sich beispielsweise Chips der Firma NVidia™, die auf der Xavier™, Volta™ oder Post-Volta™ Architektur basieren. Die Prozessoren sind unter anderem über einen Bus (angedeutet dargestellt) mit einem schnellen Speicher 23 verbunden, in denen die zu verarbeitenden Einzelbilder abgelegt sind. Darüber hinaus kann der Speicher 23 auch weitere Daten enthalten, wie beispielsweise die Vektorendefinitionen oder der Roi's und ähnliches. Mit dem schnellen Speicher 23 ist ein zweiter langsamerer Speicher 24 gekoppelt. Dieser enthält eine Videosequenz, die aus wenigstens zwei, oftmals aber mehreren Einzelbildern besteht. Die Videosequenz wird von einer Kamera 25 oder einem entsprechenden Sensor geliefert.

Figur 6 illustriert ein Ausführungsbeispiel des Verfahrens zum gemeinsamen Detektieren, Verfolgen und Klassifizieren wenigstens eines Objektes in einer Videobildsequenz. Bei dem Verfahren ist das Netzwerk

bereits trainiert, und zwar derart, dass Detektieren, Identifizieren, und Verfolgen gemeinsam trainiert wurde beispielsweise mit einem Verfahren nach dem oben dargestellten Verfahren. In einem ersten Schritt S61 werden eine Vielzahl von Bildern empfangen. Dies erfolgt vorzugsweise paarweise, d.h. es wird ein erstes und ein darauffolgendes zweites Einzelbild bereitgestellt. In einem zweiten darauffolgenden Schritt S62 werden Objekte aus dem ersten und dem zweiten Einzelbild detektiert. In einer konkreteren Ausführung werden dazu mehrere Objektvorschläge (S620) erzeugt, anhand derer die Objekte detektiert werden.

Sodann werden im Schritt S63 ein Objekt aus dem ersten und dem zweiten Einzelbild selektiert. In Schritt S64 wird ein mindestens ein Klassifizierungsvektors und eine Position für das Objekt jeweils für das erste und das zweite Einzelbild ermittelt. In Schritt S65 wird eine Korrelation zwischen den beiden ausgewählten Objekten bestimmt. Damit wird geprüft, ob die Objekte die gleichen oder verschiedenen Objekte sind. Zu diesem Zweck wird ein Assoziierungswert ermittelt, der sich aus einer Korrelationsanalyse und basierend auf dem ermittelten Klassifizierungsvektor und der Position ergibt. Dabei ist der Assoziierungswertes beispielsweise umso größer, je größer die Wahrscheinlichkeit ist, dass beide Objekte die gleichen sind.

Liegt der Assoziierungswert über einem gewissen Level oder ist er der größte Wert, sofern alle Objekte auf diese Weise miteinander verglichen wurden, so wird in Schritt S66 ein zeitlich konsistenter und eindeutiger Identifikations- oder Identifikationsvektor des mindestens einen Objektes für jedes Einzelbild in Antwort auf den ermittelten Assoziierungswert erzeugt. In diesem Zusammenhang ist eine Beziehung für die Erstellung des Assoziierungswertes herauszustellen.

Zwischen dem ersten und dem zweiten Einzelbild liegt immer eine gewisse Zeit, in der sich das Objekt von einer Position in eine andere Position bewegen kann. Die Abweichung hinsichtlich der Position kann dabei umso größer sein, je mehr Zeit zwischen dem ersten und dem zweiten Einzelbild vergeht, oder je langsamer die Bildwiederholrate

ist. Dadurch sinkt das Vertrauensniveau für eine Assoziierung basierend auf der Position oder auch der Positionsvorhersage des Objektes.

Aus diesem Grund wird für die Assoziierung nicht nur die Klassifizierung und die Position herangezogen, sondern es erfolgt auch eine

5 relative Gewichtung dieser beiden Parameter zueinander. Im Besonderen verschiebt sich die Gewichtung zugunsten des Klassifizierungsvektors je größer der Zeitabstand zwischen beiden Einzelbildern ist. Mit anderen Worten, wird die Position weniger stark gewichtet, wenn die Zeit zwischen Einzelbildern anwächst, bzw. die Bildwiederholrate sinkt. Aus
10 zwei auf diese Weise miteinander assoziierte Objekten lässt die Trajektorie des Objektes für die beiden Einzelbilder bestimmen.

Für das Detektieren eines Objektes in Schritt S62 können mehrere Aufgaben oder Zwischenschritte notwendig sein. Beispielsweise wird eine oder mehrere bounding boxen bzw. Objektvorschläge erzeugt. Die

15 bounding boxen als begrenzende Elemente umgeben dabei das mindestens eine Objekt. Neben einer Detektion der Objekte (Det_t , Det_{t-1} aus den Objektvorschlägen (Roi's) in Schritt S620 wird in Schritt S621 eine Vorhersage ($P_{t-1 \rightarrow t}$) für eine Positionsveränderung für die bounding box vom ersten Einzelbild zum zweiten Einzelbild erzeugt. Ebenso wird ein
20 Geschwindigkeitsvektor ($V_{t-1 \rightarrow t}$) für die bounding box des ersten Einzelbildes gebildet.

Für den späteren Vergleich können verschiedene Bounding Boxen und damit Objekte selektiert werden. Beispielsweise ist es möglich, eine bounding box des ersten Einzelbildes und eine bounding box des zweiten

25 Einzelbildes auszuwählen. In diesem Fall werden zwei detektierte Objekte Det_{t-1} , Det_t , die in Schritt S620 ermittelt wurden, ausgewählt (S630). Alternativ kann auch die Vorhersage $V_{t-1 \rightarrow t}$ (S620) gemeinsam mit einer bounding box und einem Objekt Det_t des zweiten Einzelbildes verwendet werden. Als dritte Möglichkeit bietet es sich an, den er-
30 mittelten Geschwindigkeitsvektors $V_{t-1 \rightarrow t}$ und Auswählen die bounding box eines Objektes des zweiten Einzelbildes zu benutzen.

Für die Bestimmung des Klassifizierungsvektors in Schritt S64 werden Merkmale des Objektes in Schritt S640 erfasst und aus den erfassten Merkmalen ein eindeutiger Merkmalsvektor bestimmt (S641). Mittels des

Merkmalsvektors kann ein Objekt klassifiziert werden, in dem der Merkmalsvektor beispielsweise mit vordefinierten Klassenvektoren verglichen wird und der Klassenvektor mit der größten Korrelation als Klasse des Objektes herangezogen wird. Beispielsweise können die Klassenvektoren aus einer vorher definierten Gruppe ausgewählt werden. Das Netzwerk, mit dem dieses Verfahren durchgeführt wird, wurde auch mit diesen Klassenvektoren trainiert.

Nach der Erzeugung des Assoziierungswertes kann für die Erzeugung eines zeitlich konsistenten und eindeutigen Identifikationsvektors ein Hungarian combinatorial optimization Verfahren zum Einsatz kommen.

Mit dem vorgeschlagenen Verfahren werden somit Objekte in zwei aufeinanderfolgenden Einzelbildern erkannt. Ihnen wird ein eindeutiger ID-Wert zugeordnet, so dass sie mit diesem auch über mehrere Einzelbilder hinweg verfolgt werden können. Insbesondere ist es durch die Erzeugung des Merkmalsvektors möglich, ein Objekt zu Re-identifizieren, d.h. es korrekt als ein bekanntes Objekt zu identifizieren auch wenn es über einige Einzelbilder hinweg nicht sichtbar ist. Das obige Verfahren kann über eine längere Videosequenz hinweg mit paarweisen Einzelbildern durchgeführt werden. Durch Verknüpfung gleicher ID-Werten über die Sequenz hinweg lassen sich Trajektorien der Objekte der Sequenz erzeugen, die für verschiedene Anwendungen, z. B. Sicherheit und Überwachung nutzbar sind.

In einigen anderen Anwendungen ist es zweckmäßig verlorengegangene Objekte wieder korrekt zu identifizieren. Ein Beispiel neben der Personenüberwachung ist auch im Bereich des autonomen Fahrens zu finden, wenn beispielsweise ein Fahrzeug während eines Überholvorgangs aus dem Gesichtsfeld der Kamera verschwindet und später wieder erscheint. So ist es nach dem Verfahren möglich, ein Objekt im zweiten Einzelbild, welches keinem Objekt im ersten Einzelbild zuordenbar ist, (beispielsweise, weil die Assoziierungswerte zu niedrig sind) mit Objekten aus vorangegangenen Einzelbildern zu vergleichen. Durch den eindeutigen ID-Wert werde so neu erscheinende Objekte mit älteren verglichen und bei Übereinstimmung re-identifiziert.

Ein anderer Aspekt der Erfindung betrifft wie oben bereits dargelegt das maschinenbasierte Training, um Objekte richtig detektieren, identifizieren und klassifizieren zu können. Während in herkömmlichen deep learning Netzwerken diese Aufgaben sequentiell trainiert werden, wird
5 erfindungsgemäß vorgeschlagen, dies parallel, gemeinsam und gleichzeitig durchzuführen. Wie bei allen Training werden hierzu eine Reihe von Testbildern vorgelegt, deren Objekte bereits korrekt klassifiziert sind. Das Training soll nun diese Objekte mit der korrekten Klassifizierung nun nicht nur detektieren und über mehrere Einzelbilder
10 verfolgen, sondern auch Merkmale erlernen, die typisch für bestimmte Klassen sind. Daneben dienen die Merkmale zur Verbesserung der Vorhersage und damit der Verfolgung.

Figur 7 zeigt eine Ausgestaltung des Verfahrens für ein maschinenbasiertes Training eines computerimplementierten Netzwerks zum gemeinsamen Detektieren, Verfolgen und Klassifizieren wenigstens eines Objektes in einer Videobildsequenz aufweisend eine Vielzahl aufeinanderfolgender Einzelbilder. Dazu wird in Schritt S71 eine Vielzahl von Einzelbildpaaren bereitgestellt. Jedes Einzelbildpaar enthält ein
15 erstes und in darauffolgendes zweites Einzelbild. Darüber hinaus weist jedes Einzelbild wenigstens zwei Objektvorschläge (RoI) aufweist, denen je ein Klassenidentifikationsvektor eindeutig zugeordnet ist. Mit anderen Worten hat jede region of interest (Roi) eine eindeutige und bekannte Klassenidentifikation. Zweckmäßig weist jeder Objektvorschlag für das Training ein Objekt auf, welches durch den Klassenidentifikationsvektor einer vordefinierten Klasse zugeordnet ist.
20
25

Mit dieser Voraussetzung beginnt das Training in Schritt S72, in dem nun für jeden der wenigstens zwei Objektvorschläge ein Merkmalsvektor ermittelt wird (S721), welche dem Klassenidentifikationsvektor zugeordnet sind. Parallel dazu wird ein Identifikationsvektor für jeden
30 Objektvorschlag bestimmt (S722). Letztlich wird eine bounding box regression für jeden Objektvorschlag bestimmt (S723), wobei diese spezifisch für den Klassenidentifikationsvektor ist.

In einem optionalen Schritt in S72 werden Merkmale einer vordefinierten Größe aus den Einzelbilder mittels eines R-CNN Verfahrens zusammengelegt (S7211) und anschließend mittels eines Faltungsblocks mit einem nachfolgenden Zusammenlegen zu einem globalen Vektor gefaltet (S7212). Dies ist zweckmäßig wenn der globale Merkmalsvektor oder auch die zusammengelegten Merkmale zur Bestimmung der bounding box regression und des Identifikationsvektors und/oder weiteren Aufgaben herangezogen wird. Dadurch wird im Gegensatz zu einer rein sequentiellen Abarbeitung Rechenzeit und Aufwand eingespart. Auf den globalen Vektor wird eine voll verbundene Schicht einer definierten Größe angewandt, um einen gegenüber dem globalen Vektor niedriger dimensionierten Identifikationsvektor zu erhalten. Dazu gehört auch ein online instance matching loss Verfahren. Durch diesen Ansatz wird ein geringerer n-dimensionaler Identitätsvektor gelernt, wobei n insbesondere 256 beträgt.

In einem optionalen Schritt S722 wird ein Klassifikationsscore durch Anwenden zweier voll verbundener Schichten der Größe $(C+1)$ und $4 \times (C+1)$ auf die Objektvorschläge ermittelt, wobei $(C+1)$ die Anzahl der Klassenidentifikationsvektoren einschließlich des Bildhintergrundes ist.

Das Verfahren fährt dann mit Schritt S73 fort und ermittelt eine für jeden der wenigstens zwei Objektvorschläge korrelierte Merkmalskarte zwischen dem ersten Einzelbild und dem darauffolgenden zweiten Einzelbild. Dazu wird eine Korrelation zwischen Merkmalen einer ersten Merkmalskarte des ersten Einzelbildes mit den Merkmalen von zu der ersten Merkmalskarte benachbarten Ausschnitten einer zweiten Merkmalskarte ermittelt. Vereinfacht gesagt, es wird eine Korrelation zwischen einem Merkmal des ersten Einzelbildes in Ausschnitten des zweiten Einzelbildes gesucht, wobei die Ausschnitte benachbart zu der korrespondierenden Position des Merkmals in dem zweiten Bild sind. Damit werden Bewegungen des Merkmals im zweiten Bild erfasst.

Diese Merkmalskarte wird in Schritt S74 dazu verwendet, eine inter-frame regression jeden Objektvorschlag zu bestimmen. Da die Ergebnisse während des Trainings bekannt sind (die für das Training vorgelegten

Bilder sind bekannt), kann nun für jeden Trainingszyklus ein Fehler ermittelt (S75) werden.

Dazu wird aber nicht der Fehler für jeden der vorangegangenen Schritte einzeln erfasst, sondern ein kombinatorischer Fehler. Dieser ergibt
5 sich aus den Fehlern des Bestimmens des Klassenidentifikationsvektors, des Bestimmens des wenigstens eines Identifikationsvektors, des Bestimmens der spezifischen bounding box regression und des Bestimmens der inter-Frame regression. In einem Fall kann der kombinatorische Fehler durch die Summe aus diesen Fehlern gebildet werden (S75). Für
10 das Training wird dieser kombinatorische Fehler nun minimiert.

In einem Beispiel wird für die Bestimmung des Fehlers bei der Klassifizierung ein Softmax Cross entropy loss Verfahren angewandt (S751). Ein Smooth L1-loss Verfahren kann zur Bestimmung des bounding box regression Fehlers benutzt werden wie in Schritt S752 durchgeführt.
15 Das gleiche Verfahren wird auch in Schritt S753 eingesetzt, um den Fehler einer inter-Frame regression zu ermitteln.

Grundsätzlich sind die in dieser Anmeldung offenbarte Verfahren nicht nur auf Videosequenzen oder Bildern im sichtbaren Bereich, sondern auch auf sensorische Informationen in anderen Spektralbereichen oder
20 sogar andere sensorische Informationen, beispielsweise Schall oder Radar übertragbar.

Bezugszeichenliste

	1, 25	Kamera
	2	Netzwerk
	3	Verfolgungsmodul
5	4	Anwendung
	5	Vorhersagemodul
	6	Identifizierungsmodul
	7	Detektionsmodul
	20	Computer
10	21, 22	Prozessoren
	23	Speicher
	24	Videozwichenspeicher
	31	Transporter, Objekt
	32	Bus, Objekt
15	33	PKW, Objekt
	34	PKW, Objekt
	DetIdenNet	Detektions- und Identifizierungsmodul
	PredNet	Vorhersagemodul
	BaseNet	Basismodul
20	f^{t-1} , f^t	Merkmale

Ansprüche

1. Verfahren für ein maschinenbasiertes Training eines computerimplementierten Netzwerks zum gemeinsamen Detektieren, Verfolgen und Klassifizieren wenigstens eines Objektes in einer Videobildsequenz aufweisend eine Vielzahl aufeinanderfolgender Einzelbilder, das
5 Verfahren aufweisend die Schritte:

- Bereitstellen (S71) einer Vielzahl von Einzelbildpaaren, wobei jedes Einzelbildpaar ein erstes und in darauffolgendes zweites Einzelbild umfasst, wobei jedes Einzelbild wenigstens einen Objektvorschlag (RoI) aufweist, denen je ein Klassenidentifikationsvektor eindeutig zugeordnet ist;
10
- für jeden Objektvorschlag (S72):
 - o Bestimmen (S721) eines Merkmalsvektors, welche dem Klassenidentifikationsvektor zugeordnet sind;
 - 15 o Bestimmen (S722) eines Identifikationsvektors;
 - o Bestimmen (S723) einer Klassenidentifikationsvektor spezifischen bounding box regression;
- Bestimmen (S73) einer korrelierten Merkmalskarte zwischen dem ersten Einzelbild und dem darauffolgenden zweiten Einzelbild;
- 20 - Bestimmen (S75) für jeden Objektvorschlag eine inter-frame regression;
- Minimieren (S76) eines kombinierten Fehlers, der sich aus den Fehlern des Bestimmens des Klassenidentifikationsvektors, des Bestimmens des Identifikationsvektors, des Bestimmens der spezifischen bounding box regression und des Bestimmens der inter-
25 Frame regression ergibt.

2. Verfahren nach Anspruch 1, umfassend den Schritt für den wenigstensn einen Objektvorschlag:

- 30 - Zusammenlegen von Merkmalen einer vordefinierten Größe aus den Einzelbilder mittels eines ROI-Align Verfahrens; und
- Falten der zusammengelegten Merkmale mittels eines Faltungsblocks mit einem nachfolgenden Zusammenlegen zu einem globalen Vektor.

3. Verfahren nach Anspruch 1, bei dem das Bestimmen eines Identifikationsvektors umfasst:
- Verwenden eines online instance matching loss, um einen n-dimensionalen Identitätsvektor zu lernen, wobei n insbesondere 256 beträgt.
4. Verfahren nach Anspruch 2, bei dem das Bestimmen des Identifikationsvektors umfasst:
- Anwenden einer voll verbundenen Schicht einer definierten Größe, um einen gegenüber dem globalen Vektor niedriger dimensionierten Identifikationsvektor zu erhalten.
5. Verfahren nach einem der Ansprüche 1 bis 4, weiter umfassend für jeden des wenigstens einen Objektvorschlag den Schritt eines:
- Ermitteln eines Klassifikationsscores durch Anwenden zweier voll verbundener Schichten der Größe $(C+1)$ und $4 \times (C+1)$, wobei $(C+1)$ die Anzahl der Klassenidentifikationsvektoren einschließlich des Bildhintergrundes ist.
6. Verfahren nach einem der Ansprüche 1 bis 5, bei dem der Schritt des Minimierens eines kombinierten Fehlers umfasst:
- Anwenden eines Softmax Cross entropy loss Verfahrens zur Bestimmung des Fehlers bei der Klassifizierung; und/oder
 - Anwenden eines Smooth L1-loss Verfahrens zur Bestimmung des bounding box regression Fehlers; und/oder
 - Anwenden Smooth L1-loss Verfahrens zur Bestimmung des Fehlers einer inter-Frame regression.
7. Verfahren nach einem der Ansprüche 1 bis 6, bei dem der Schritt des Bestimmens einer korrelierten Merkmalskarte umfasst:
- Ermitteln einer Korrelation zwischen der ersten Merkmalskarte des ersten Einzelbildes benachbarten Ausschnitten in der Merkmalskarte des zweiten Einzelbildes.
8. Verfahren nach einem der Ansprüche 1 bis 7, wobei der wenigstens eine Objektvorschlag ein Objekt aufweist, welches durch den

Klassenidentifikationsvektor einer vordefinierten Klasse zugeordnet ist.

9. Verfahren zum gemeinsamen Detektieren, Verfolgen und Klassifizieren wenigstens eines Objektes in einer Videobildsequenz aufweisend eine Vielzahl aufeinanderfolgender Einzelbilder mit einem trainierten computerimplementierten Netzwerk, insbesondere mit einem nach dem Verfahren der vorherigen Ansprüche trainierten Netzwerks, umfassend die Schritte:

- Empfangen (S61) eines ersten Einzelbildes und eines darauffolgenden zweiten Einzelbildes;
- Detektieren (S62) mindestens eines Objektes in dem ersten Einzelbild und dem darauffolgenden zweiten Einzelbild;
- Auswählen (S63) eines Objektes aus dem ersten und dem zweiten Einzelbild;
- Ermitteln (S64) mindestens eines Klassifizierungsvektors und einer Position für das Objekt aus dem ersten und dem zweiten Einzelbild;
- Ermitteln (S65) eines Assoziierungswertes basierend auf dem ermittelten Klassifizierungsvektor und der Position;
- Erzeugen (S66) eines zeitlich konsistenten und eindeutigen Identifikationsvektors des mindestens einen Objektes für jedes Einzelbild in Antwort auf den ermittelten Assoziierungswert; wobei für die Erzeugung des Assoziierungswerts eine von der Zeit zwischen dem ersten und zweiten Einzelbild abhängigen relative Gewichtung zwischen dem ermittelten Klassifizierungsvektor und der Position vorgesehen ist.

10. Verfahren nach Anspruch 9, bei dem der Schritt eines Detektierens mindestens eines Objektes umfasst:

- Erzeugen (S620) einer bounding box, welche das mindestens eine Objekt umgibt;
- Erzeugen (S621) einer Vorhersage für die bounding box vom ersten Einzelbild zum zweiten Einzelbild; und
- Erzeugen (S622) eines Geschwindigkeitsvektors für die bounding box des ersten Einzelbildes.

11. Verfahren nach Anspruch 10, bei dem eine bounding box für jedes des mindestens einen Objektes vorgesehen ist.
12. Verfahren nach einem der Ansprüche 10 bis 11, bei dem der
5 Schritt des Auswählens wenigstens einen der folgenden Schritte umfasst:
- Auswählen (S630) der bounding box des ersten Einzelbildes und Auswählen der bounding box des zweiten Einzelbildes;
 - Auswählen (S631) der Vorhersage und Auswählen der bounding box
10 des zweiten Einzelbildes; und
 - Auswählen (S632) des Geschwindigkeitsvektors und Auswählen der bounding box des zweiten Einzelbildes.
13. Verfahren nach einem der Ansprüche 9 bis 12, bei dem der
15 Schritt des Ermitteln mindestens eines Klassifizierungsvektors umfasst:
- Erfassen (S640) von Merkmalen des Objektes;
 - Berechnen (S641) eines eindeutigen Merkmalsvektors aus den erfassten Merkmalen; und
 - Klassifizieren (S642) des Objektes aus einer Gruppe vorgegebener
20 Klassen anhand der erfassten Merkmale oder anhand des berechneten Merkmalsvektors.
14. Verfahren einem der Ansprüche 9 bis 12, bei dem mit steigender Zeit (oder einer fallenden Bildrate) zwischen dem ersten und zweiten Einzelbild die relative Gewichtung zwischen dem ermittelten Klassifizierungsvektor und der Position ansteigt.
15. Verfahren einem der Ansprüche 9 bis 12, bei dem für den
30 Schritt des Erzeugens einer zeitlich konsistenten und eindeutigen Identifikationsvektor ein Hungarian combinatorial optimization Verfahren eingesetzt wird.
16. Verfahren nach einem der Ansprüche 9 bis 15, bei dem die
35 eindeutige Identifikation eines Objekts eines zweiten Einzelbildes, dass keinem Objekt eines ersten Einzelbildes zuordenbar

ist mit der Identifikation eines Objektes eines dritten Einzelbildes verglichen wird, welches zeitlich vor dem ersten Einzelbild liegt.

- 5 17. Ein System zur Klassifizierung von Objekten auf einem Computer, der umfasst:
- einen Speicher
 - einen oder mehrere Prozessoren, die ausgeführt sind, das Verfahren nach einem der Schritte 9 bis 16 auszuführen.
- 10 18. Ein auf einem Medium abgespeichertes Computerprogrammprodukt mit Instruktionen, die auf einem oder mehreren Prozessoren ausgeführt, das Verfahren nach einem der Schritte 9 bis 16 durchführen.
- 15 19. Anordnung, vorzugsweise eine computerimplementierte Anordnung mit
- einem Zwischenspeicher für wenigstens ein erstes und ein zweites Einzelbild einer Videosequenz;
 - 20 - ein RPN-Modul zur Erzeugung einer ersten Vielzahl unabhängiger Objektvorschläge aus dem ersten Einzelbild und einer zweiten Vielzahl unabhängiger Objektvorschläge aus dem zweiten Einzelbild;
 - ein mit einem Vorhersagemodul gemeinsam trainiertes Bestimmungs- und Identifikationsmodul, welches mit dem RPN-Modul gekoppelt ist und ausgeführt ist, Objekte aus der ersten Vielzahl und der zweiten Vielzahl von Objektvorschlägen zu identifizieren und im ersten und zweiten Einzelbild zu detektieren und welches weiter ausgeführt ist, aus den Objektvorschlägen eine Vorhersage der Position eines Objektes im ersten Einzelbild und im zweiten Einzelbild zu treffen;
 - 25 - das mit dem Bestimmungs- und Identifikationsmodul gemeinsam trainierte Vorhersagemodul, welches mit dem RPN-Modul gekoppelt ist und ausgeführt ist, eine inter-Frame Vorhersage für jedes
 - 30 Objekt im ersten Einzelbild basierend auf der ersten Vielzahl
 - 35

von Objektvorschlägen und der zweiten Vielzahl von Objektvorschlägen zu treffen.

- 5 20. Anordnung nach Anspruch 19, bei dem dem RPN-Modul eine BaseNet Architektur vorgeschaltet ist, die ausgeführt ist, eine Vielzahl von Basismerkmalen für das erste und das zweite Einzelbild zu ermitteln, welche dem RPN-Modul zur Erzeugung der Objektvorschläge zuführbar sind.
- 10 21. Anordnung nach einem der Ansprüche 19 bis 20, bei dem das trainierte Bestimmungs- und Identifikationsmodul ausgeführt ist, für jeden Vorschlag Merkmale einer vorbestimmten Größe aus den Basismerkmalen zusammenzufassen, um einen mehrdimensionalen Merkmalsvektor bereitzustellen.
- 15 22. Anordnung nach Anspruch 21, bei dem das trainierte Bestimmungs- und Identifikationsmodul ausgeführt ist, mit Hilfe des Merkmalsvektors einen Klassifizierungsparameter des Objektes im Objektvorschlag zu bestimmen.
- 20 23. Anordnung nach Anspruch 21, bei dem der Klassifizierungsparameter aus einer Menge vorbestimmter Klassen ausgewählt ist.
- 25 24. Anordnung nach einem der Ansprüche 19 bis 23, bei dem das trainierte Bestimmungs- und Identifikationsmodul ausgeführt ist, ein Objekt in dem Objektvorschlag einen eindeutigen Identifikationsvektor zuzuordnen.
- 30 25. Anordnung nach einem der Ansprüche 19 bis 24, bei dem das trainierte Bestimmungs- und Identifikationsmodul ausgeführt ist, für jeden Objektvorschlag eine bounding box regression zu erzeugen.
- 35 26. Anordnung nach einem der Ansprüche 20 bis 25, bei dem das trainierte Vorhersagemodul ausgeführt ist, anhand der Basismerkmale des ersten und zweiten Einzelbildes die Regressionsziele

der detektierten Objekte vom ersten zum zweiten Einzelbild vorherzusagen.

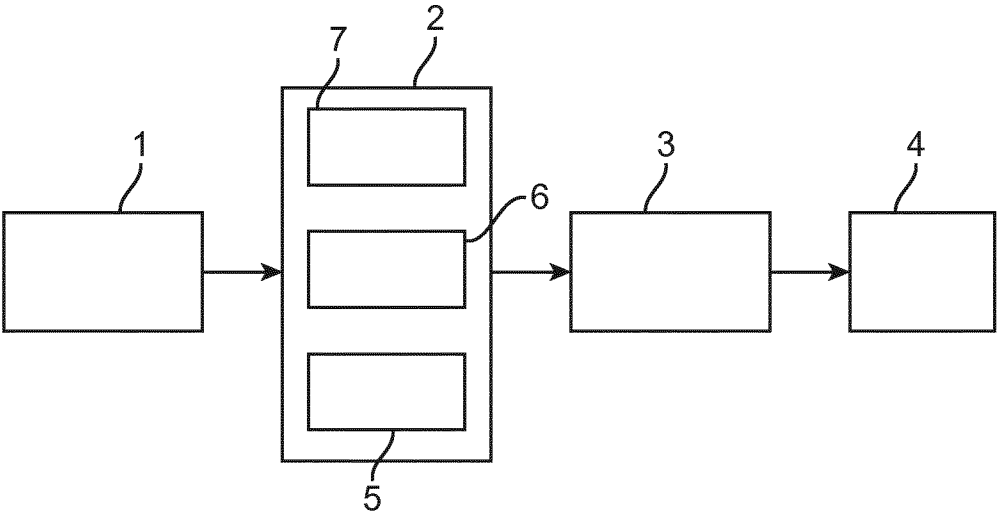


FIG. 1

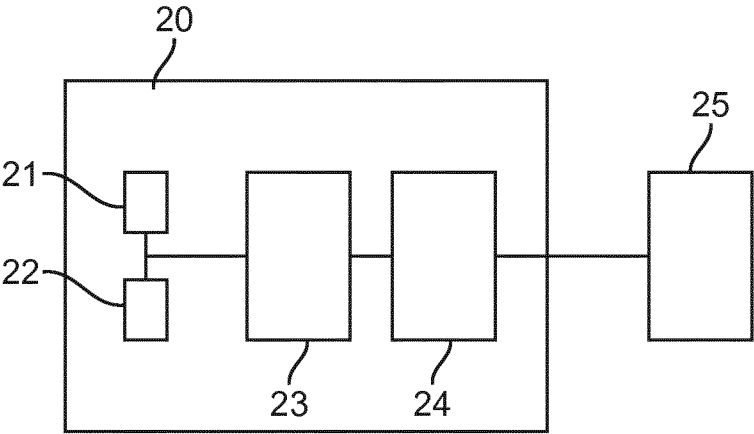


FIG. 2

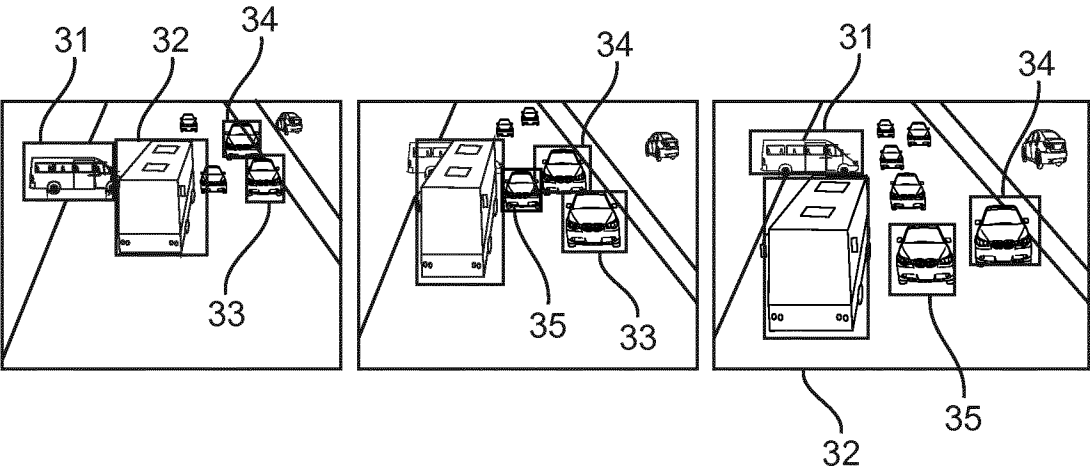


FIG. 3

2 / 4

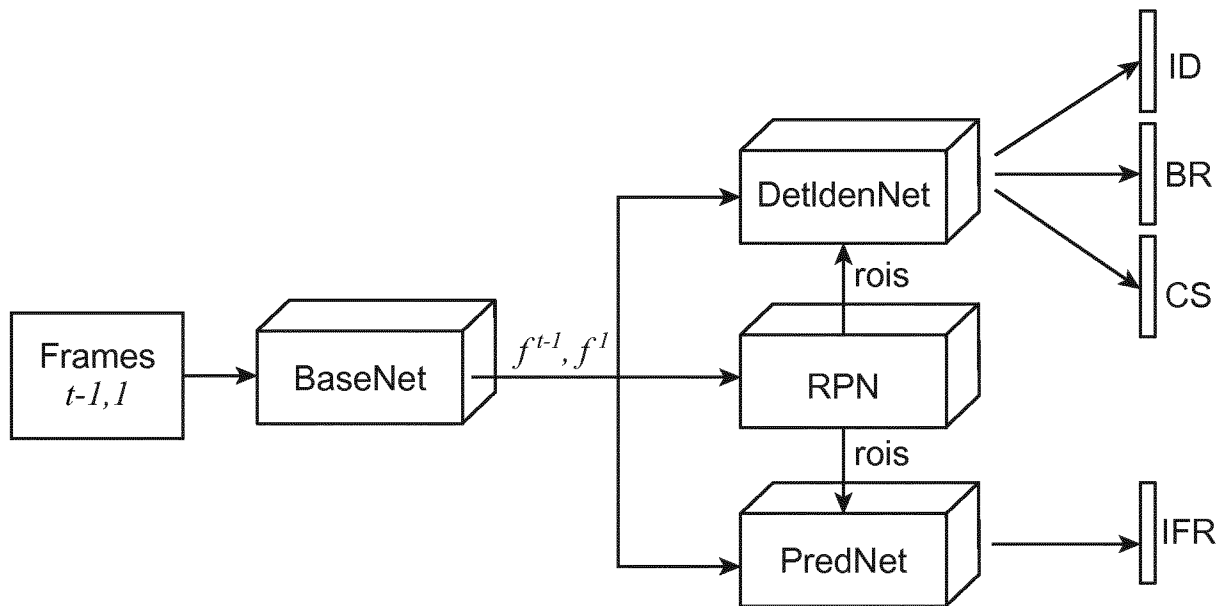


FIG. 4

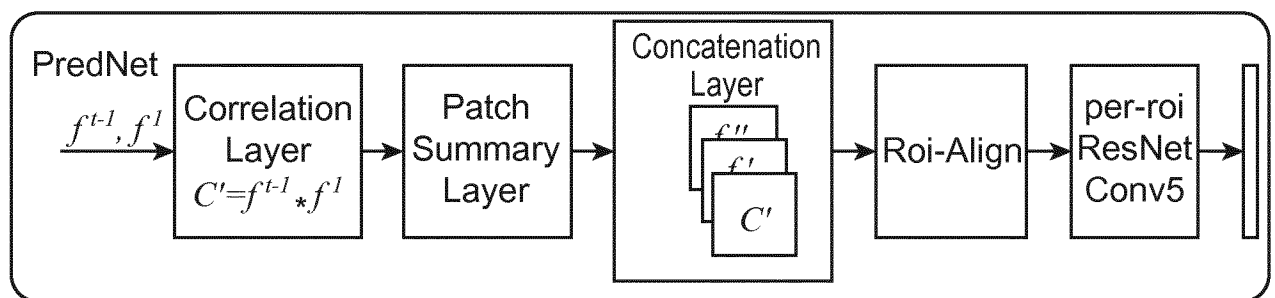


FIG. 5

3 / 4

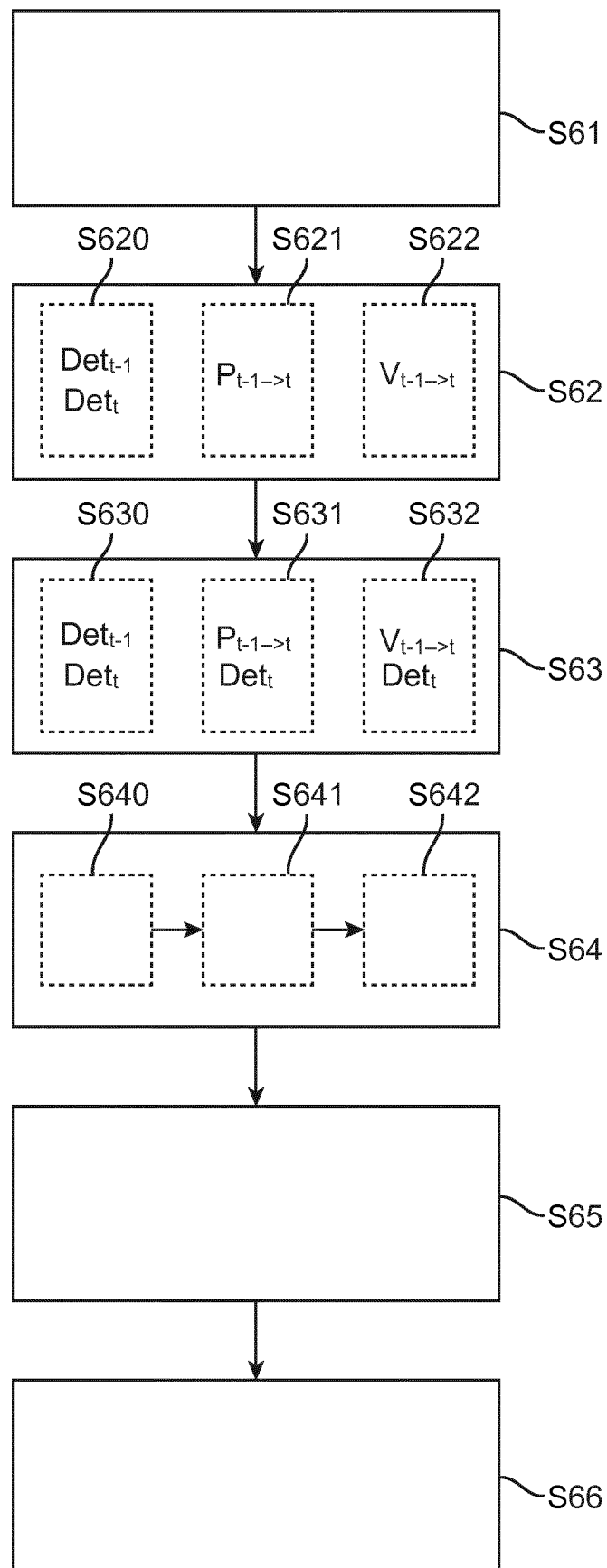


FIG. 6

4 / 4

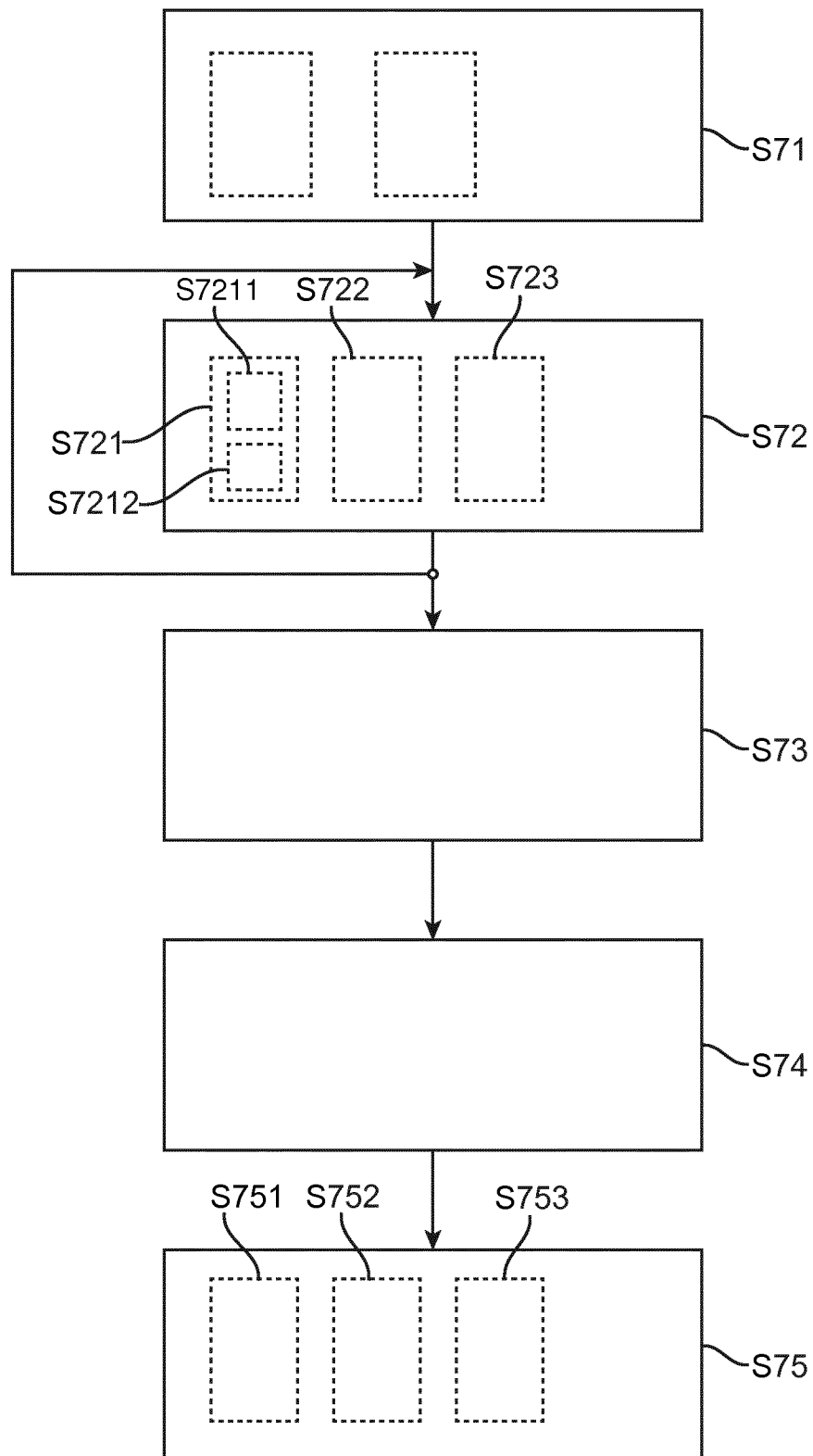


FIG. 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/EP2019/081317

A. CLASSIFICATION OF SUBJECT MATTER**G06K 9/00**(2006.01)i; **G06K 9/32**(2006.01)i; **G06K 9/62**(2006.01)i; **G06T 7/246**(2017.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06K; G06T

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CHRISTOPH FEICHTENHOFER ET AL. "Detect to Track and Track to Detect" <i>ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853</i> , 11 October 2017 (2017-10-11), XP081310251 sections 2 and 3; figure 2	1-26
A	XIAO TONG ET AL. "Joint Detection and Identification Feature Learning for Person Search" <i>IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. PROCEEDINGS, IEEE COMPUTER SOCIETY, US</i> , 21 July 2017 (2017-07-21), pages 3376-3385, [retrieved on 2017-11-06] DOI: 10.1109/CVPR.2017.360 ISSN: 1063-6919, XP033249686 section 3; figures 2 and 3	1-26



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

13 February 2020

Date of mailing of the international search report

20 February 2020

Name and mailing address of the ISA/EP

European Patent Office
p.b. 5818, Patentlaan 2, 2280 HV Rijswijk
Netherlands

Telephone No. (+31-70)340-2040

Facsimile No. (+31-70)340-3016

Authorized officer

Atmatzidis, Lazaros

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/EP2019/081317

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	LUO WENJIE ET AL. "Fast and Furious: Real Time End-to-End 3D Detection, Tracking and Motion Forecasting with a Single Convolutional Net" 2018 <i>IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, IEEE</i>, 18 June 2018 (2018-06-18), pages 3569-3577 DOI: 10.1109/CVPR.2018.00376 XP033476327 section 4	1-26

A. KLASSIFIZIERUNG DES ANMELDUNGSGEGENSTANDES		
INV. G06K9/00	G06K9/32	G06K9/62 G06T7/246
ADD.		
Nach der Internationalen Patentklassifikation (IPC) oder nach der nationalen Klassifikation und der IPC		
B. RECHERCHIERTE GEBIETE		
Recherchierter Mindestprüfstoff (Klassifikationssystem und Klassifikationssymbole)		
G06K G06T		
Recherchierte, aber nicht zum Mindestprüfstoff gehörende Veröffentlichungen, soweit diese unter die recherchierten Gebiete fallen		
Während der internationalen Recherche konsultierte elektronische Datenbank (Name der Datenbank und evtl. verwendete Suchbegriffe)		
EPO-Internal		
C. ALS WESENTLICH ANGESEHENE UNTERLAGEN		
Kategorie*	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
X	CHRISTOPH FEICHTENHOFER ET AL: "Detect to Track and Track to Detect", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 11. Oktober 2017 (2017-10-11), XP081310251, Abschnitt 2 und 3; Abbildung 2 ----- -/-	1-26
☒ Weitere Veröffentlichungen sind der Fortsetzung von Feld C zu entnehmen ☐ Siehe Anhang Patentfamilie		
* Besondere Kategorien von angegebenen Veröffentlichungen : "A" Veröffentlichung, die den allgemeinen Stand der Technik definiert, aber nicht als besonders bedeutsam anzusehen ist "E" frühere Anmeldung oder Patent, die bzw. das jedoch erst am oder nach dem internationalen Anmeldedatum veröffentlicht worden ist "L" Veröffentlichung, die geeignet ist, einen Prioritätsanspruch zweifelhaft erscheinen zu lassen, oder durch die das Veröffentlichungsdatum einer anderen im Recherchenbericht genannten Veröffentlichung belegt werden soll oder die aus einem anderen besonderen Grund angegeben ist (wie ausgeführt) "O" Veröffentlichung, die sich auf eine mündliche Offenbarung, eine Benutzung, eine Ausstellung oder andere Maßnahmen bezieht "P" Veröffentlichung, die vor dem internationalen Anmeldedatum, aber nach dem beanspruchten Prioritätsdatum veröffentlicht worden ist "T" Spätere Veröffentlichung, die nach dem internationalen Anmeldedatum oder dem Prioritätsdatum veröffentlicht worden ist und mit der Anmeldung nicht kollidiert, sondern nur zum Verständnis des der Erfindung zugrundeliegenden Prinzips oder der ihr zugrundeliegenden Theorie angegeben ist "X" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann allein aufgrund dieser Veröffentlichung nicht als neu oder auf erfinderischer Tätigkeit beruhend betrachtet werden "Y" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese Verbindung für einen Fachmann naheliegend ist "&" Veröffentlichung, die Mitglied derselben Patentfamilie ist		
Datum des Abschlusses der internationalen Recherche		Absendedatum des internationalen Recherchenberichts
13. Februar 2020		20/02/2020
Name und Postanschrift der Internationalen Recherchenbehörde Europäisches Patentamt, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016		Bevollmächtigter Bediensteter Atmatzidis, Lazaros

C. (Fortsetzung) ALS WESENTLICH ANGESEHENE UNTERLAGEN		
Kategorie*	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
A	XIAO TONG ET AL: "Joint Detection and Identification Feature Learning for Person Search", IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. PROCEEDINGS, IEEE COMPUTER SOCIETY, US, 21. Juli 2017 (2017-07-21), Seiten 3376-3385, XP033249686, ISSN: 1063-6919, DOI: 10.1109/CVPR.2017.360 [gefunden am 2017-11-06] section 3; Abbildungen 2,3 -----	1-26
A	LUO WENJIE ET AL: "Fast and Furious: Real Time End-to-End 3D Detection, Tracking and Motion Forecasting with a Single Convolutional Net", 2018 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, IEEE, 18. Juni 2018 (2018-06-18), Seiten 3569-3577, XP033476327, DOI: 10.1109/CVPR.2018.00376 [gefunden am 2018-12-14] section 4 -----	1-26